JBC RESEARCH ARTICLE

The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function

Received for publication, January 14, 2022, and in revised form, September 6, 2022 Published, Papers in Press, September 12, 2022,
https://doi.org/10.1016/j.jbc.2022.102480

Leesa J. Klau[1,2,‡], Sheila Podell[1,‡], Kaitlin E. Creamer[1,‡], Alyssa M. Demko[1], Hans W. Singh[1], Eric E. Allen[1,3], Bradley S. Moore[1,4], Nadine Ziemert[1], Anne Catrin Letzel[1], and Paul R. Jensen[1,*]

From the [1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA; [2]Department of Biotechnology and Food Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway; [3]Molecular Biology Section, Division of Biological Sciences, University of California San Diego, La Jolla, California, USA; [4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA

Edited by F. Peter Guengerich

The Natural Product Domain Seeker (NaPDoS) webtool detects and classifies ketosynthase (KS) and condensation domains from genomic, metagenomic, and amplicon sequence data. Unlike other tools, a phylogeny-based classification scheme is used to make broader predictions about the polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) genes in which these domains are found. NaPDoS is particularly useful for the analysis of incomplete biosynthetic genes or gene clusters, as are often observed in poorly assembled genomes and metagenomes, or when loci are not clustered, as in eukaryotic genomes. To help support the growing interest in sequence-based analyses of natural product biosynthetic diversity, here we introduce version 2 of the webtool, NaPDoS2, available at http://napdos.ucsd.edu/napdos2. This update includes the addition of 1417 KS sequences, representing a major expansion of the taxonomic and functional diversity represented in the webtool database. The phylogeny-based KS classification scheme now recognizes 41 class and subclass assignments, including new type II PKS subclasses. Workflow modifications accelerate run times, allowing larger datasets to be analyzed. In addition, default parameters were established using statistical validation tests to maximize KS detection and classification accuracy while minimizing false positives. We further demonstrate the applications of NaPDoS2 to assess PKS biosynthetic potential using genomic, metagenomic, and PCR amplicon datasets. These examples illustrate how NaPDoS2 can be used to predict biosynthetic potential and detect genes involved in the biosynthesis of specific structure classes or new biosynthetic mechanisms.

Increased access to DNA sequence data coupled with a better understanding of the molecular genetics of natural product biosynthesis are driving major advances in natural products research. These advances have been facilitated by webtools such as antiSMASH 6.0 (1) and PRISM 4 (2) that identify natural product biosynthetic gene clusters (BGCs) from assembled sequence data and provide insight into the types of small molecules produced (3). These tools have proven instrumental for genome mining research (4), while others have been developed to address more specific topics such as resistance-guided antibiotic discovery (5), biosynthetic gene biogeography (6), and *trans*-acyl transferase (*trans*-AT) substrate specificity (7). The Natural Product Domain Seeker (NaPDoS) is a specialized webtool used to assess biosynthetic diversity based on short sequence tags and thus does not require BGC assembly (8). It targets ketosynthase (KS) and condensation (C) domains to make broader predictions about the polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes, respectively, in which they reside. NaPDoS detects these domains and classifies them using a phylogeny-based scheme that reflects well-supported biosynthetic knowledge and established PKS and NRPS function.

Here, we introduce version 2 of the NaPDoS webtool (NaPDoS2), which includes an updated KS database and classification scheme that better reflects the expanded taxonomic distributions and functional diversity of PKSs. These enzymes produce structurally diverse polyketides ranging from lipids to macrolides, which represent an important source of compounds for pharmaceutical and other biotechnological applications (9). Polyketides also serve important ecological functions ranging from antioxidants (10) to allelochemicals (11) and thus can provide insight into how organisms interact with each other and the environment. PKSs share much in common with fatty acid synthases (FASs), generating

‡ These authors contributed equally to this work.
* For correspondence: Paul R Jensen
Present addresses for Alyssa M Demko: Smithsonian Marine Station, Fort Pierce, FL 34949, United States.
Present addresses for Nadine Ziemert: Interfaculty Institute of Microbiology and Infection Medicine, Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Auf der Morgenstelle 28, 72,076 Tübingen, Germany. German Centre for Infection Research (DZIF), Partner Site Tübingen, Germany.

compounds *via* the successive decarboxylative condensation and processing of acyl-CoA precursors (12). PKSs have been broadly divided into three types based on their organization and function (13) of which NaPDoS2 detects and classifies KSs associated with types I and II. While canonical type I PKSs have a modular organization and function in an assembly line fashion, some function iteratively while others (*e.g.*, *trans*-AT) lack a cognate acyl-transferase domain (9, 14). Similarly, canonical type II PKSs function iteratively and were originally best known to produce aromatic polyketides. Yet some type II PKSs function noniteratively (13), while others produce linear specialized metabolites (15). A central feature of PKSs is the KS domain, which in most cases catalyzes a Claisen condensation between the extender unit and the growing, thioester-linked polyketide chain. In recent years, our knowledge of KS functional diversity has expanded significantly to reveal new enzymology and diverse product outcomes across biology. The broad distributions and functional specificities among type I and II PKSs can, in many cases, be resolved phylogenetically (16–18), with these evolutionary relationships forming the basis of the NaPDoS2 classification scheme.

The NaPDoS2 website, available at http://napdos.ucsd.edu/napdos2/, includes many updated features that improve the usability of the tool for natural product discovery. Here, we report database and pipeline modifications that provide broader taxonomic coverage, better resolution among functionally characterized PKSs, a new subclassification scheme for type II PKSs, an increased capacity for processing large datasets, and the enhanced detection of eukaryotic KSs. Statistical validation tests have been used to select parameters for optimizing sensitivity and specificity, including both detection and classification accuracy based on query sequence length. The upgraded webtool was used to demonstrate the utility of NaPDoS2 for predicting biosynthetic potential in genomic, metagenomic, and amplicon datasets.

Results and discussion

Pipeline efficiency and interface upgrades

As in the original release (8), NaPDoS2 detects and classifies KS and C domains from nucleotide or amino acid sequence data. While both versions follow the same general workflow, substitution of the more computationally efficient program DIAMOND (19) for NCBI BLAST (20), eliminates the need for an extra Hidden Markov Model prefiltering step (Fig. S1). Speed improvements using DIAMOND are relatively modest for small jobs but can reach orders of magnitude for large datasets, especially those consisting of short query sequences (19). The NaPDoS2 pipeline executes most rapidly on amino acid sequences, which do not need to be translated. Although processing times increase with total nucleotide sequence length and the number of database matches, results for microbial genomes, assembled metagenomes, and PCR KS amplicons containing thousands of hits can typically be obtained within seconds to minutes (Table S1). These improvements enable users to perform large-scale analyses that were not feasible with the original NaPDoS release.

User interface upgrades include the addition of a "Domain Classification Summary" page that lists the total number of domains detected in the query data as grouped by their NaPDoS2 classification (Fig. S2*A*). This page provides the option to select classes of specific interest for more detailed investigation (Fig. S2*B*), which is particularly useful when large numbers of domains are detected. Each BGC represented in the database is linked to a summary page that includes a representative structure and the classification of each KS or C domain within that BGC (Fig. S2*C*). The independent classification of each domain is particularly useful given that a single gene or BGC can contain multiple domain types (21). New webtool features also include quick start instructions outlining the NaPDoS2 workflow (Fig. S3*A*) and a downloadable documentation and tutorial file.

Database expansion

The primary goals for NaPDoS2 were to expand the KS database and classification scheme to include biosynthetic functions that were not represented in the original release, to supplement poorly populated classes, and to provide greater taxonomic coverage. One thousand four hundred seventeen new KS sequences were added, raising the database total to 1877 (average length 418 ± 49 amino acids), an increase of 308% (Fig. S4). Most of the additional sequences were derived from the MIBiG repository of experimentally verified BGCs (22), including new type II PKSs, type I fungal PKSs, and type I FASs from both bacterial and fungal sources. A few uncharacterized protist and metazoan sequences were included due to the scarcity of experimental verification among these groups. Although 84 C domain sequences were added, their classification scheme has not been updated and remains an important goal for future releases. The taxonomic breakdown of the current NaPDoS2 KS database sequences is 93.8% bacteria, 4.7% fungi, and 1.5% other eukaryotes (Table S2), reflecting the taxonomic skew of available experimental data. To improve result interpretation, database (match) IDs now display the BGC name, domain number, class, and subclass identifiers for each domain.

Phylogeny-based KS classification

The ability to predict PKS and NRPS biosynthetic potential based on KS or C domains distinguishes NaPDoS2 from other bioinformatic tools. The domain classification scheme is derived from KS sequence phylogenies and their relationships to established biochemical functions, gene architectures, and structural features of the compounds produced. While not all classes and subclasses are monophyletic, functionally coherent clades form the basis of the classification scheme. Phylogenies generated from the updated KS database allowed us to establish 41 class and subclass assignments associated with type I and II PKSs and FASs (Table S2 and Fig. S5). This represents an increase of 410% over the original release. To verify the sufficiency of using KS domains to indicate broader PKS context, we assessed the genomic neighborhoods of KSs detected in a variety of genomic and metagenomic datasets. In

all cases, the NaPDoS2 KS classification agreed with the antiSMASH 6.0 (1) classification (Fig. S6). In some cases, NaPDoS2 provided more detailed information including the identification of type II subclasses, omega-3 polyunsaturated fatty acid (PUFA) KSs, and highly, partially, and nonreducing fungal KSs. NaPDoS2 can also distinguish among different KS classes within a single gene or BGC (Fig. S6).

With the classification scheme established, a simplified reference tree was generated using 414 sequences representing all class and subclass assignments (Fig. 1). Broadly, this reference phylogeny delineates the well-established relationships between FASs and type I and II PKSs (17). The enhanced classification of type I FAS KSs from bacteria/fungi, protists, and metazoa allows users to better distinguish between KSs associated with specialized metabolism and fatty acid biosynthesis across diverse taxa. The expanded eukaryotic type I coverage includes KSs from fungal iterative *cis*-AT PKSs and several protist (Phyla Amoebozoa and Apicomplexa) and metazoan (Phyla Chordata, Echinodermata, and Nematoda) PKSs, including those linked to characterized natural products from *Caenorhabditis elegans* (23) and *Dictyostelium discoideum* (24). The seven type I PKS classes described in the original NaPDoS release have been reorganized into three classes (modular *cis*-AT, iterative *cis*-AT, and *trans*-AT) and 14 subclasses in the NaPDoS2 release. Example structures associated with each KS type are shown in Figure 2 with associated metadata (Table S3) and descriptions of each KS type (Fig. S5 and website documentation) provided to help connect KSs with structures and relevant references.

Modular cis-AT KSs

The modular *cis*-AT class comprises KSs associated with the canonical assembly line PKSs (*e.g.*, erythromycin biosynthesis) and now includes four subclasses of which two [olefin synthase and tandem enoyl-CoA hydratase (ECH)] are new to NaPDoS2 (Fig. 1). The olefin synthase subclass is best known from cyanobacteria and is associated with the formation of a terminal olefin on a fatty acyl precursor (25). These KS sequences form two clades in the reference tree, which reflects the

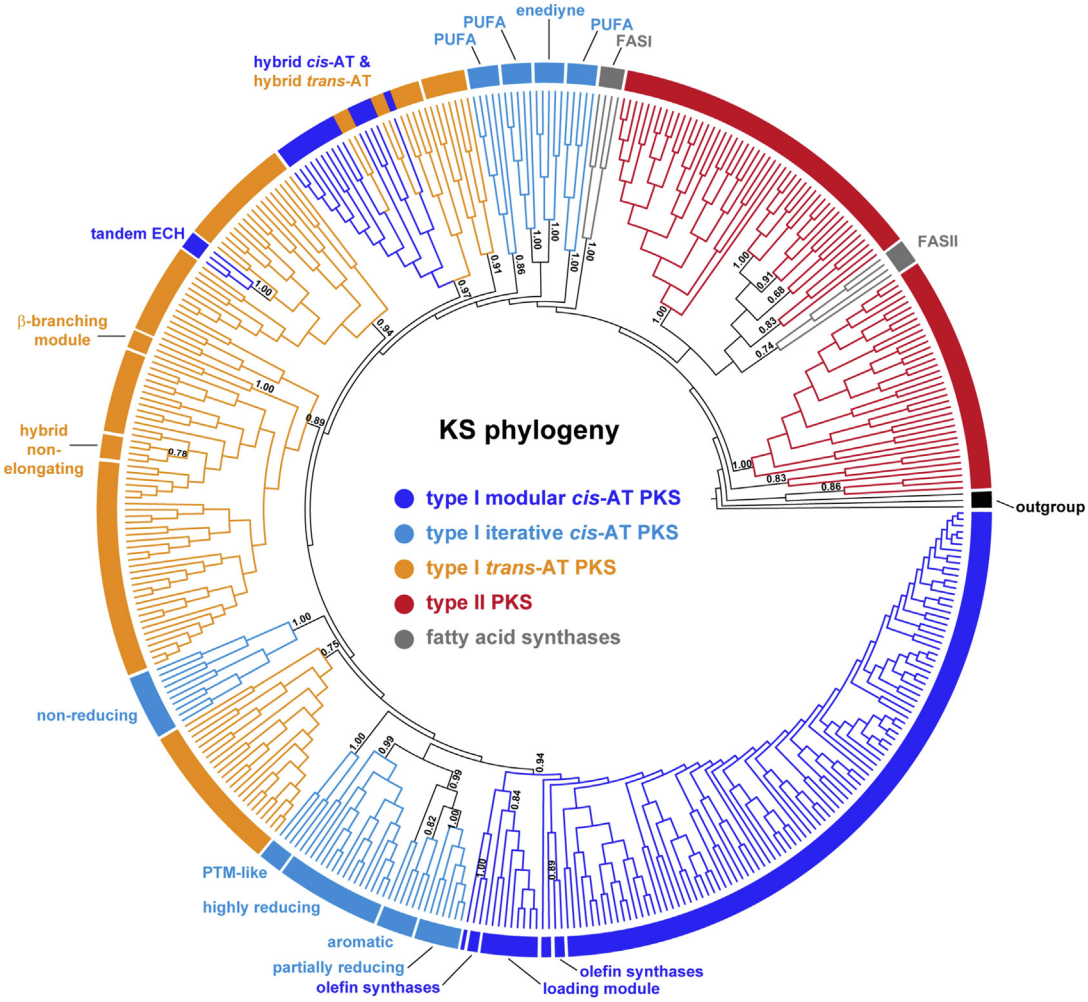

Figure 1. KS phylogeny-based classification. Maximum likelihood phylogeny generated from 414 KS sequences. Clades are color-coded and labeled according to their NaPDoS2 classification. Transfer bootstrap expectation (TBE) support was estimated using Booster (72). The full name of each sequence can be viewed in Fig. S7, which can be used to link a query match to a specific location in the tree. Thiolases from *Escherichia coli*, *Zoogloea ramigera*, and *Streptomyces avermitilis* were used as outgroups. An expanded phylogeny of the type II KSs is presented in Figure 3. FAS, fatty acid synthase; KS, ketosynthase; NaPDoS2, Natural Product Domain Seeker version 2; PKS, polyketide synthase; PUFA, polyunsaturated fatty acid.

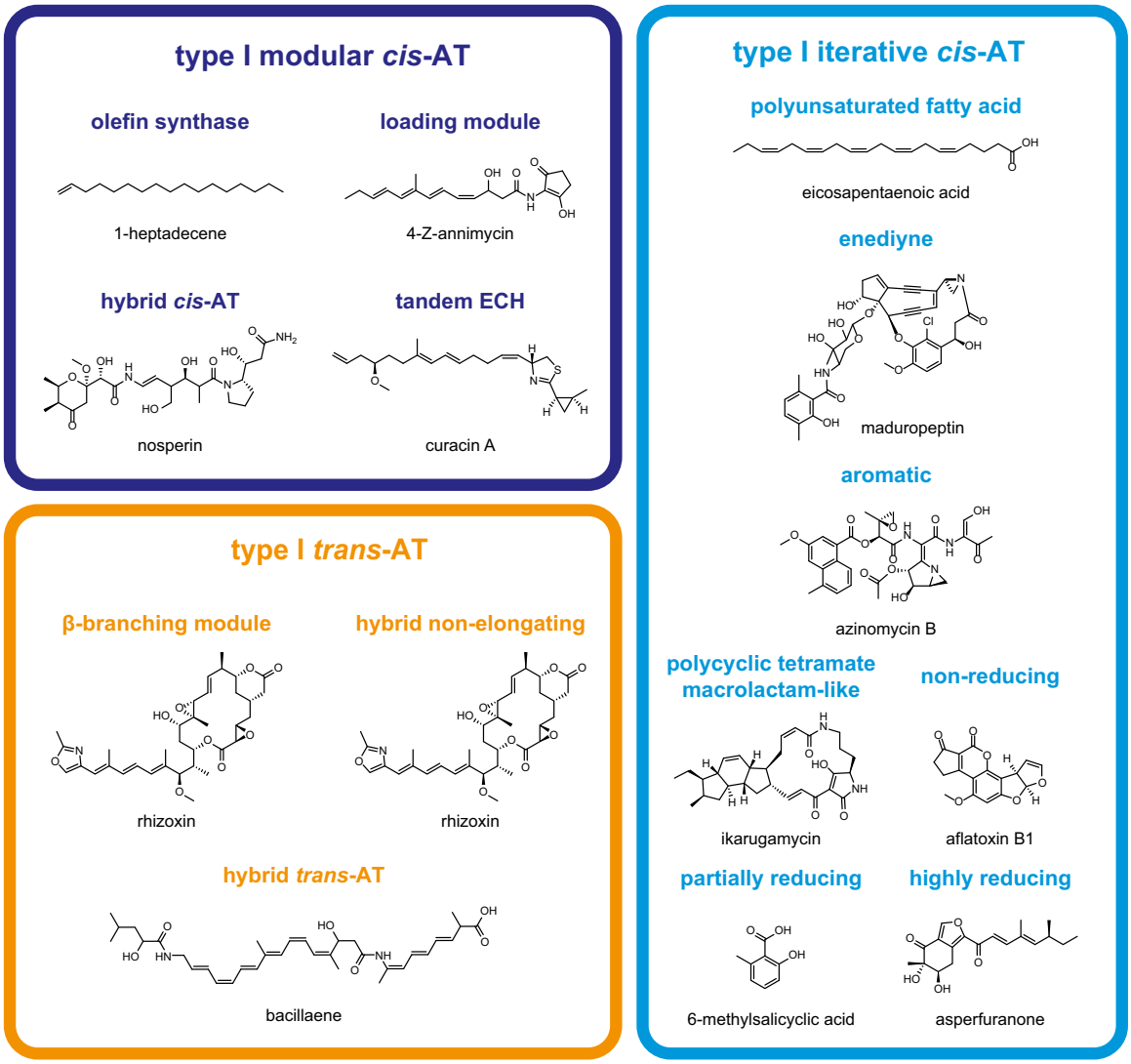

Figure 2. Example structures for the type I KS classes and subclasses recognized by NaPDoS2. Descriptions of each KS type, associated metadata, and references can be found in Fig. S5 and Table S3. Information on each structure can also be accessed from the BGC tab on the website (note: 1-heptadecene = *Moorea producens* olefin synthase, 4-Z-annimycin = annimycin, and eicosapentaenoic acid = *Schizochytrium* polyunsaturated fatty acid on the website). Colors correspond to Figure 1. KS, ketosynthase; NaPDoS2, Natural Product Domain Seeker version 2.

sporadic distribution of the OLS pathway among cyanobacteria (25). The tandem ECH subclass is associated with gene cassettes that introduce a branch at the β-keto position. In these PKSs, the KS domain is located immediately downstream of tandem ECH and enoyl reductase (ER) domains, which catalyze decarboxylation and reduction reactions, respectively, as seen in cylindrocyclophane biosynthesis (26) and reviewed elsewhere in detail (27). Most of the modular *cis*-AT KSs in the database are not associated with a specialized function and thus are not assigned to a subclass.

Iterative cis*-AT KSs (bacteria)*

The iterative *cis*-AT class (iPKSs) maintains a modular organization, with multiple enzymatic domains on a single protein, yet instead of functioning as an assembly line these enzymes catalyze an iterative series of elongation steps (28). The expanded NaPDoS2 classification scheme now includes seven iPKS subclasses. The four observed in bacteria include the aromatic and polycyclic tetramate macrolactam (PTM) subclasses (29), both of which are new to NaPDoS2, and the enediyne and PUFA subclasses, which were identified in the original release. PUFA KSs now form three clades in the reference tree, which correlate with the three KS domains typically present in PUFA PKSs (30). Aromatic iPKSs generally produce simple monocyclic or bicyclic aromatic compounds that are distinct from enediynes and PUFAs (29). The PTM-like iPKSs produce complex compounds containing a macrocyclic lactam with an embedded tetramic acid moiety fused with a polycyclic system (*i.e.*, 2–3 rings) derived from polyene chains (29). The ability to quickly recognize PTM biosynthetic potential provides opportunities to expand on the number of compounds discovered in this unusual and often biologically active class (31). New iterative type I PKSs continue to be

discovered, including some that are unexpectedly abundant and widespread in the genus *Streptomyces* (31), providing opportunities to expand this class in future updates beyond the seven subclasses currently recognized by NaPDoS2.

Iterative cis*-AT KSs (fungi)*

Iterative *cis*-AT KSs are also observed in fungi and can be delineated into highly reducing (HR), partially reducing, and NR iPKS subclasses (32). These subclasses differ in the number and type of β-keto processing ketoreductase, dehydratase, and ER domains present. HR iPKS BGCs, which contain all three β-keto processing domains, often contain a C-methylation domain responsible for α-carbon methylation and some are fused with NRPS modules (33). Their products are usually linear or cyclic, nonaromatic compounds. Partially reducing iPKS BGCs lack ER and sometimes also dehydratase domains. Due to their similarity to bacterial aromatic iPKSs, they have been hypothesized to originate from a horizontal gene transfer event between bacteria and fungi (34). NR iPKS BGCs lack all three β-keto processing domains and have specialized domains that facilitate starter unit loading and product folding (33).

Trans-AT KSs

KS domains from *trans*-AT PKSs form multiple clades in the reference tree (Fig. 1), which may reflect KS substrate-specificity (35). Nonetheless, three functionally distinct subclasses can be recognized by NaPDoS2 of which those associated with β-branching modules and hybrid, nonelongating KSs are new to NaPDoS2. The β-branching module, comprising a KS domain, a cryptic "B" domain, and an acyl carrier protein (ACP) domain, represents one of the mechanisms for the formation of β-branching in *trans*-AT PKSs (36, 37). Hybrid, nonelongating KSs (KS^0) typically follow an NRPS module, lack the catalytic histidine required for elongation, and are proposed to facilitate the transfer of the growing polyketide to the next module (7, 38). The hybrid *trans*-AT KS subclass clades with the hybrid *cis*-AT subclass and functions similarly in modules that lack AT domains.

Type II KSs

A major aim for the NaPDoS2 upgrade was to expand the classification of type II KS domains based on established functions and evolutionary relationships (39). A phylogeny of the type II KS domains in the updated database (Fig. 3) provided sufficient resolution to delineate five functionally defined classes and four sequences that have yet to be assigned a functional class (annotated as type II unclassified). This represents a major improvement over the original NaPDoS release, which simply identified a sequence as being associated with a type II PKS. The most highly populated of the five type II KS classes recognized by NaPDoS2 are associated with PKSs that function iteratively to produce reactive poly-β-ketoacyl intermediates that cyclize to polycyclic aromatic compounds. The type II aromatic class could be further delineated into nine functionally defined subclasses based on a concatenated alignment of the two KS subunits, which generated a better

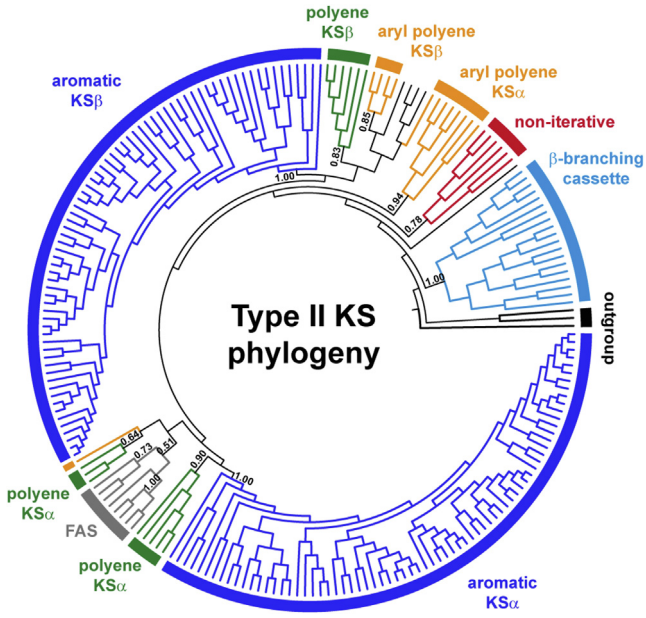

Figure 3. Type II KS phylogeny-based classification. Maximum likelihood phylogeny of 212 KS sequences. Clades are color-coded and labeled according to their NaPDoS2 classification. TBE bootstrap support was estimated using Booster (72). See Fig. S8 for sequence annotations. Thiolases from *E. coli*, *Z. ramigera*, and *S. avermitilis* were used as outgroups. Type II KS sequences that lack further annotation have yet to be assigned a functional class. KS, ketosynthase; NaPDoS2, Natural Product Domain Seeker version 2; TBE, transfer bootstrap expectation.

resolved phylogeny (Fig. 4). Example structures (Fig. 5) along with the associated metadata (Table S3) help connect each KS type with structures and relevant references.

Type II KSs function as a heterodimer, with the KSα subunit responsible for the iterative elongation of a nascent poly-β-keto chain, the KSβ subunit determining the chain length (also referred to as the chain-length factor), and both subunits mediating the first ring cyclization reaction (38, 40). The KSα and KSβ subunits are implemented as individual sequences in the NaPDoS2 workflow thus allowing them to be distinguished. To define these subclasses, the tree topology was compared with poly-β-keto chain length, carbon-carbon position of the first ring cyclization, type of starter unit, and the core and intermediate structures of the products (Fig. 4). These subclasses are consistent with previous phylogenies (16, 41) and chemotypes identified in PKMiner (42) and antiSMASH version 5.0 (43) based on full-length BGC analyses. KSs from several unusual type II aromatic PKSs that are poorly represented in the database (*e.g.*, resistomycin biosynthesis) were not assigned subclasses and are instead annotated as "type II aromatic unclassified" in NaPDoS2.

The four remaining type II KS classes are more unusual and include one that is responsible for introducing β-branching in *trans*-AT PKSs. These KSs are associated with HMGS or HCS (3-hydroxy-3-methylglutaryl coenzyme A synthase) gene cassettes that encode several monodomain proteins including a decarboxylating KS that can be resolved in the type II KS phylogeny (Fig. 3) (27). Another unusual type II KS class recognized in the phylogeny functions noniteratively. The KSs associated with these enzymes facilitate the incorporation of

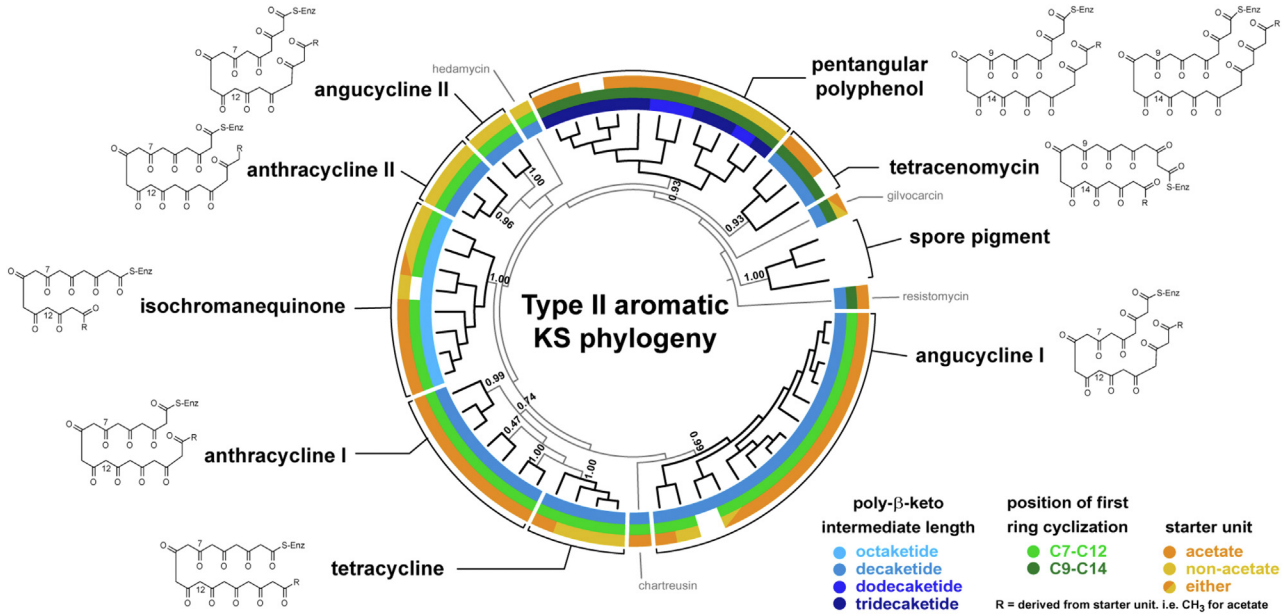

Figure 4. Type II aromatic KS phylogeny-based classification. Maximum likelihood phylogeny of concatenated KSα and β subunits from 59 type II BGCs. Clades are annotated based on NaPDoS2 classification. Structural motifs associated with subclasses shown in colored rings (white, not determined). See Fig. S9 for sequence annotations and Table S3 for biosynthetic references. TBE bootstrap support was estimated using Booster (72). BGCs, biosynthetic gene clusters; KS, ketosynthase; NaPDoS2, Natural Product Domain Seeker version 2; TBE, transfer bootstrap expectation.

acetate, propionate, and succinate building blocks leading to the production of macrocyclic compounds such as the nonactin antibiotics, in which case they catalyze C-O bond formation (44). Noniterative type II PKSs contain multiple type II and III KSs in which each domain is responsible for a single condensation step (45, 46). The remaining type II classes represent KSs associated with the production of highly reduced polyenes or aryl polyenes (Fig. 3), with the latter representing the most prominent biosynthetic family observed across a wide taxonomic distribution of bacterial genomes (47). Similar to type II aromatic PKSs, polyene and aryl polyene PKS also contain KSα and KSβ subunits with biosynthesis proceeding through a series of alternating elongation and reduction steps (15, 48, 49). Aryl polyene PKSs also possess homodimeric KSs that function to complement the heterodimer (15, 49). These are most similar to KSs observed in type II FASs (not shown in the tree) and are classified as such in the NaPDoS2 database.

Performance evaluation

Recognizing that computational prediction algorithms are seldom perfect, leave-one-out cross validation (50) and receiver operating characteristic (ROC) curves (51) are powerful statistical tools for quantifying sensitivity and specificity but require calibration using known positive and negative control datasets that maximize discriminatory power. A set of positive controls (213 sequences) for evaluating NaPDoS2 performance was obtained by clustering all full-length, experimentally verified KS domains in the NaPDoS2 database at 50% amino acid identity. These sequences belong to five conserved domain families within the condensing enzyme superfamily of the NCBI Conserved Domain Database (Fig. S10, *A* and *B*), which encompasses enzymes that catalyze decarboxylating or nondecarboxylating Claisen-like condensation reactions for the synthesis and degradation of fatty acids and polyketides from all kingdoms of life (23% eukaryota, 70% bacteria, and 7% archaea). Negative controls (308 sequences) were selected from condensing enzyme families falling immediately outside the NaPDoS2 clades and similarly clustered at 50% amino acid identity (Fig. S10*A*). These negative controls, which include beta-ketoacyl-ACP synthases, ketoacyl-ACP synthases III, type III chalcone and stilbene synthases, thiolases, and sterol carrier protein-associated thiolases, were augmented with additional sequences (52) (Fig. S10*C*) and KSs from type III PKSs retrieved from MIBiG 2.0 (Fig. S10*D*).

Leave-one-out cross-validation scores were generated for positive and negative control sequences based on the e-values of their closest nonself match using both the original NaPDoS and NaPDoS2 KS reference databases. ROC curves were calculated from these scores to generate area under the curve (AUC) values (Fig. 6*A*), demonstrating the superior performance of NaPDoS2 (AUC = 0.987) *versus* the original release (AUC = 0.978). ROC curves were further used to establish e-value cutoff points that maximize sensitivity and minimize false positives, identifying optimum values as 1e-8 for NaPDoS2 and 4e-11 for the original NaPDoS release. Although these cutoff values provided equivalent sensitivity (detection of true positives) for their respective algorithms, the highest achievable specificity obtainable for NaPDoS was 93.0% (7% false positive rate) *versus* 97.2% for NaPDoS2 (2.8% false positive rate). This improvement is most likely explained by the expanded database underlying the NaPDoS2 pipeline. Consequences of these statistically identified differences in real-world use cases are presented in the "NaPDoS2 applications" section below.

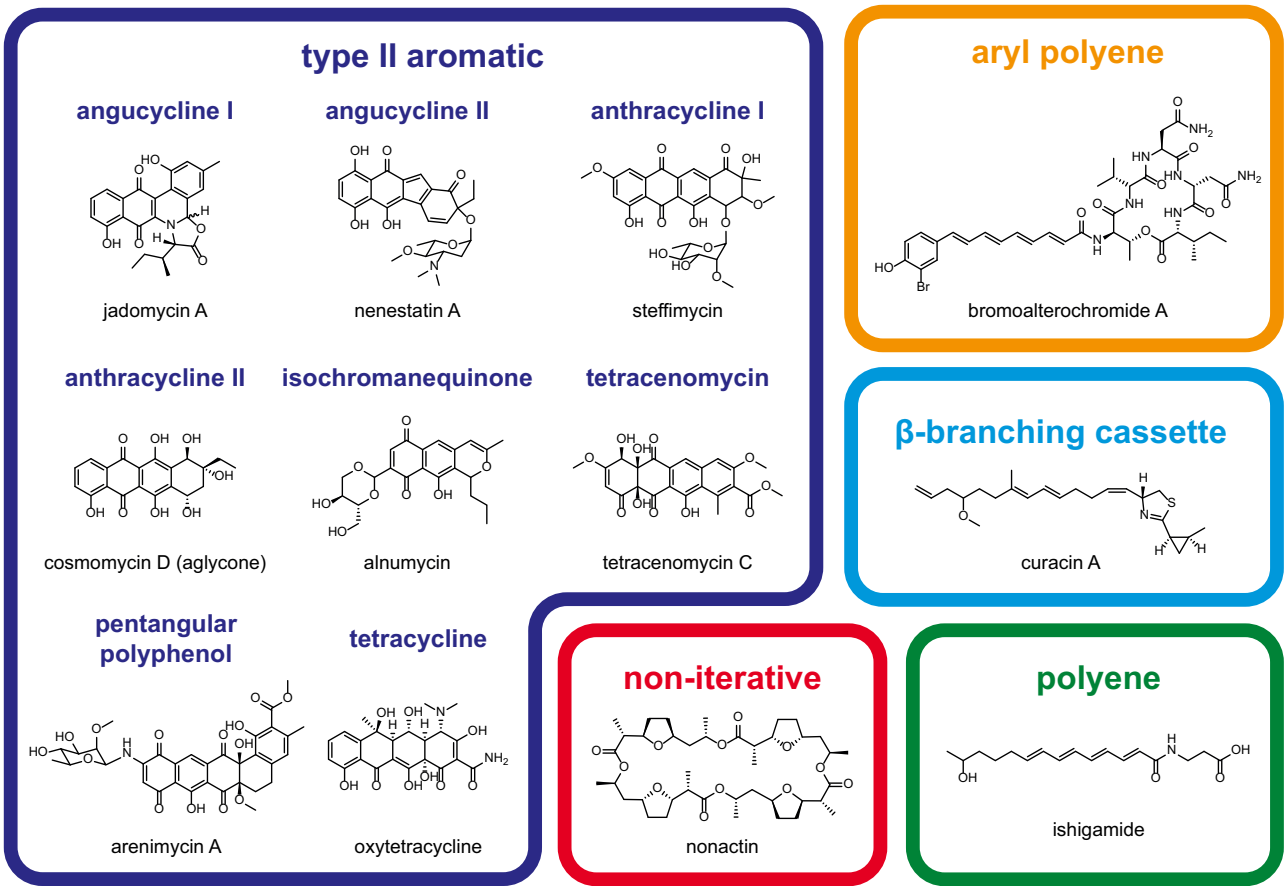

Figure 5. Example structures for the type II KS classes and subclasses recognized by NaPDoS2. Descriptions of each KS type, associated metadata, and references can be found in Fig. S5 and Table S3. Information on each structure can also be accessed from the BGC tab on the website (note: bromoalterochromide A = alterochromide on the website). Colors correspond to Figure 3. BGC, biosynthetic gene cluster; KS, ketosynthase; NaPDoS2, Natural Product Domain Seeker version 2.

The effects of partial KS sequences on detection and classification accuracy were evaluated using both full-length database sequences (typically 425 amino acids) and shorter overlapping subsets of 30, 50, 100, and 200 amino acids (aa) covering the entire length of these sequences. These subsets were designed to mimic domain fragments encountered in draft genomes, metagenomic assemblies, or KS amplicon sequences. Using the previously established 1e-8 cutoff for maximum sensitivity, NaPDoS2 detected 99% of the 200aa length subsequences as KS domains, of which 85% were correctly classified (Fig. 7). Detection and classification accuracy declined with shorter sequences, falling dramatically for sequences <50 amino acids. Length-dependent performance degradation was also observed using 1e-5 and 1e-10 e-values as cutoff scores (Fig. S11). These results illustrate the difficulty of analyzing unassembled, next-generation sequencing reads and short contigs covering partial domain fragments.

Based on performance evaluation results, default NaPDoS2 parameters were set at an e-value of 1e-8 and a minimum alignment length of 200 amino acids; however, users may choose to adjust these settings for their individual datasets. Sensitivity declines for partial KS domains can be partially offset by decreasing BLAST stringency, at the cost of increasing false positives and misclassifications. PCR-generated amplicons can provide better sensitivity than similarly sized random sequence fragments but may still be too short to obtain accurate classifications. Confident assignments to poorly populated classes or subclasses are particularly challenging given the sequence diversity observed within highly populated classes and subclasses (Fig. S12). Some ambiguities may be resolved by using NaPDoS2 sequence alignments to generate detailed phylogenetic trees, but others will remain until additional functional studies are reported.

NaPDoS2 applications

Large-scale performance evaluations targeting genomic, metagenomic, and amplicon sequence data from a variety of bacterial, fungal, metazoa, and environmental sources were conducted to demonstrate the utility of NaPDoS2 in identifying polyketide biosynthetic potential (Table 1; accession numbers in Table S3). These examples include the integration of NaPDoS2 output with other webtools (Fig. S3*B*).

Genomes

While the original NaPDoS release simply identified KSs as type II, the new release provides more sensitive detection and more detailed classifications. We analyzed 118 *Salinispora*

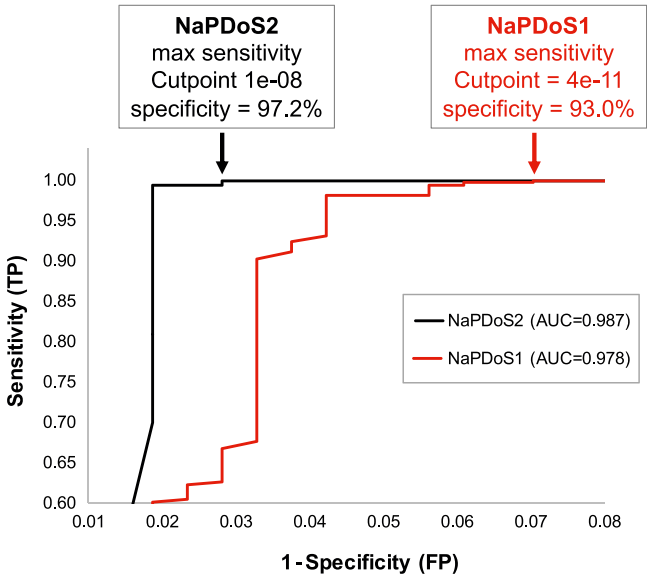

Figure 6. Receiver operating characteristic curves for NaPDoS and NaPDoS2. Nonredundant sets of 213 positive and 308 negative control KS domains were compared based on the BLASTP e-values of their closest non-self-match in each NaPDoS database. Optimal cutoff values were selected to maximize sensitivity. FP, false positives; KS, ketosynthase; NaPDoS2, Natural Product Domain Seeker version 2; TP, true positives.

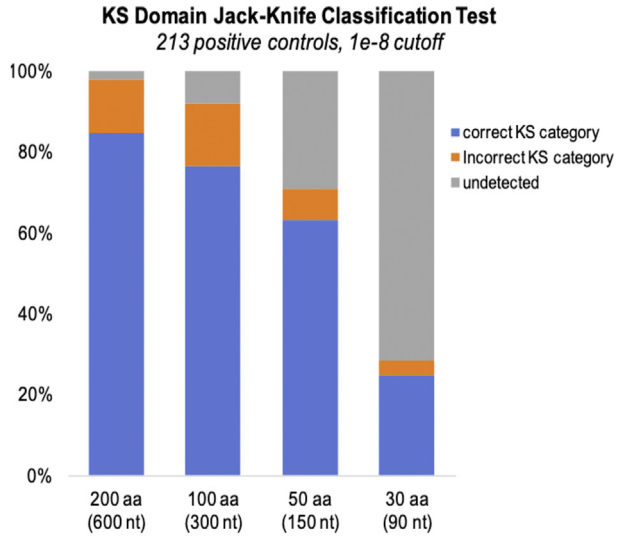

Figure 7. Effect of query size on KS detection and classification accuracy. Classifications were based on a 1e-8 BLASTP e-value cutoff score for the closest non-self-database match. Test sequences of varying lengths were obtained as overlapping sliding window subsequences covering the full length of 213 nonredundant, positive control KS domains. KS, ketosynthase.

genomes (53) with NaPDoS2 and detected a total of 662 type II KS domains in contrast to 363 using the original release (Table S4). The type II KSs detected by NaPDoS2 were delineated into seven functionally defined classes and subclasses; those that were unassigned may represent new functional diversity. A broader summary of all KSs detected in these genomes provided the first evidence that *S. arenicola* has the potential to produce PTMs (Table S5), a class of structurally complex natural products that exhibits diverse biological activities (54). These results provide new insight into the biosynthetic potential of this marine actinomycete genus.

Another important NaPDoS2 improvement is the ability to detect and classify eukaryotic KS sequences. This is illustrated by the analysis of 27 taxonomically diverse fungal genomes (55), where KSs ranged from one in *Malassezia globosa* to 50 in *Aspergillus niger* (Table S6) and the majority could be assigned to the HR subclass. We next analyzed all 159 fungal PKS BGCs in the MIBiG 2.0 repository (22) and identified 182 KS domains, all of which matched MIBiG 2.0 descriptions and literature reports (Table S7). In contrast, the original NaPDoS release only identified 14 KS domains from these same BGCs. Although relatively few metazoan PKSs have been experimentally characterized, NaPDoS2 recovered the recently described FAS and FAS-like KSs from the *Elysia chlorotica* sacoglossan genome (Table S8) (56, 57) and correctly classified them as metazoan type I FASs. A phylogenetic tree generated using NaPDoS2 confirmed their divergence from previously characterized animal FASs (data not shown). In an exploratory search for KSs in other eukaryotic genomes, NaPDoS2 detected 37 modular *cis*-AT domains, 17 type I FAS domains, 2 type II FAS domains, and 1 protist-type KS domain from the dinoflagellate *Symbodium minimus* (58), and six type II FAS domains and one type II KSα aromatic anthracycline domain from the diatom *Nitzschia inconpicua* (59), thus further validating its utility for analyzing eukaryotic sequences. While these data show that NaPDoS2 can detect and classify KS domains from complex metazoan datasets, it is best suited for predicted proteins, translated coding sequences, or transcriptomes, since it cannot excise introns associated with eukaryotic sequence data.

Metagenomes

A single NaPDoS2 analysis can provide a simultaneous overview of both bacterial and eukaryotic polyketide and fatty acid biosynthetic potential in large metagenomic datasets. While it is not reliant on fully assembled BGCs, assembled contigs are recommended to avoid the reduced classification accuracies associated with short sequence reads (Figs. 7 and S11). To illustrate this application, we assessed 20 assembled marine sediment metagenomes deposited in the Paired Omics Data Platform (60). We observed a wide range in the numbers and types of KSs detected, which can provide important insight when selecting samples for further study (Table S9). These results show how NaPDoS2 can be used to identify samples with the potential to produce compounds in rare but biologically active classes, such as enediynes and PTMs, to expand on poorly understood metazoan PKS diversity and, in cases with low sequence similarity to database or BLAST matches, detect new functional diversity. Trimmed domains can be used as search queries to assess broader genomic context (Fig. S6), compare with previously reported BGCs (22), and potentially identify the host organism when phylogenetic markers are encountered in the KS-containing contig.

Amplicons

NaPDoS2 is particularly useful for the analysis of KS/C domain PCR amplicon datasets, where it can be used not only

Table 1
NaPDoS2 applications

Table #	Application	Data type	Biological source	Dataset	Ref.
S4, S5	Bacterial type II KS	Genome	Bacteria	118 *Salinispora* strains	(53)
S6	Fungal KS & FAS	Genome	Fungi	27 Fungal spp.	(55)
S7	Fungal KS & FAS	Genome	Fungi	159 Fungal MIBiG 2.0 PKS BGCs	(22)
S7	Environmental type II KS	Amplicon	Environmental DNA	147 KS clones from soil	(61)
S8	Eukaryotic KS & FAS	Genome	Metazoa	*Elysia chlorotica*	(56, 57)
S9	Environmental type I & II KS, FAS	Metagenome	Environmental DNA	20 Marine sediment samples	(60)
S10	Environmental type I KS	Amplicon	Environmental DNA	eSNaPD v2.0 KS sequences from soil	(62)
S11	Environmental type II KS	Amplicon	Environmental DNA	Type II KS sequences from 12 soil samples	(63)
S12	Environmental and cultured type I & type II KS	Amplicon	Bacteria, environmental DNA	Type I and II KS sequences from lake sediment and enrichment cultures	(64)

The utility of NaPDoS2 was demonstrated across a variety of data types and biological sources. Details for each analysis can be found in Tables S4–S12 as summarized below. Table S3 lists accession numbers for all analyses.

to classify sequences into specific functional categories and assign a top BGC product match but also to assess primer specificity and remove nontarget sequences prior to downstream analysis. We illustrate the applications of NaPDoS2 using four KS amplicon datasets, starting with 147 KS sequences cloned from soil eDNA using type II–specific KS primers (61). Both versions of NaPDoS identified all 147 KSs as type II, while NaPDoS2 further delineated them into three type II aromatic subclasses, which agrees with the original report (Table S7). We next compared the NaPDoS2 output with that from eSNaPD v 2.0, which relates KS amplicon sequences to a database of characterized BGCs (6, 62). NaPDoS2 detected all 381 KS sequences in the eSNaPD v 2.0 New Mexico desert soil library "NM_KS_ARRY_LIB01" dataset (62) and classified the vast majority as *cis*-AT modular (Table S10). Additionally, NaPDoS2 classified virtually all KS sequences within what eSNaPD2 v 2.0 listed as novel clusters, providing new information about the biosynthetic diversity within this library (6, 62) (Table S10).

Larger KS amplicon datasets generated from soil eDNA and lake sediment enrichment cultures using type I and II specific primers were also evaluated (63, 64). While the original analyses estimated biosynthetic potential based solely on the closest MIBiG repository match, NaPDoS2 further delineated the sequences into specific type I and type II classes and subclasses. Analysis of the type II soil amplicons (63) revealed that many were not identified as KSs and appear to represent off-target sequences. Of those identified as KSs, a number were classified by NaPDos2 as type II FASs and a few as type I PKSs (Table S11). This illustrates the application of NaPDoS2 to assess primer specificity and identify potential nontarget KS sequences prior to downstream analyses.

Finally, we addressed the question of whether lowering the minimum alignment length for amplicons might increase the number of false positives, as previously observed for random domain fragments (Fig. 7). To do this, we analyzed 40,000 randomly selected sequences from longer amplicon (602 bp) type I and II KS datasets from lake sediment enrichment cultures (64) using a range of minimum amino acid alignment lengths (5000 of these sequences shown in Table S12). We placed both hit and nonhit sequences in a phylogenetic context within the condensing enzyme superfamily tree (52, 64) and mapped the conserved domains with TREND (65) (Fig. S13). Sequences identified by NaPDoS2 as KSs, regardless of alignment length (30–200aa), fall within the two clades associated with the NaPDoS2 database positive control sequences. Conversely, the sequences that NaPDoS2 did not identify as KSs fell outside of these clades and were associated with off-target, nonketosynthase domains such as AMP-binding domains and multiple phage-related domains. These results confirm that shorter NaPDoS2 alignment length settings can be used to assess polyketide biosynthetic potential from amplicon datasets and highlight the value of verifying detection and classification accuracy using phylogenetic approaches and conserved domain architectures.

Conclusions

While several tools can detect and classify the gene clusters associated with natural product biosynthesis (1, 6), NaPDoS2 employs KS and C domains as sequence tags to predict biosynthetic potential. This approach makes it well-suited for nonclustered or incomplete BGCs, amplicons, and eukaryotic genomes where other tools are less effective. The NaPDoS2 update features an expanded KS database and classification scheme that better reflects the broader taxonomic and functional diversity now recognized among type I and II PKSs. It provides a single workflow to detect and classify KS domains from diverse biological origins including bacteria, fungi, and other eukaryotes and to distinguish among those involved in fatty acid and specialized metabolite production. Updates to workflow efficiency can now accommodate the larger genomic, metagenomic, and amplicon datasets achievable with next-generation sequencing. NaPDoS2 provides a rapid method to identify microorganisms or environments with the potential to yield rare classes of compounds, such as those produced by noniterative type II PKSs (13). It can be used to prioritize samples for cultivation and to identify potentially new biosynthetic mechanisms when sequences are phylogenetically distinct from those previously characterized, as recently demonstrated for the standalone KS (*salC*) that functions as an aldolase/β-lactone synthase in salinosporamide A biosynthesis (66). The NaPDoS2 upgrades expand on the PKS diversity that can be detected with this tool and provide a method to quickly assess biosynthetic potential in a

manner that facilitates more targeted approaches to natural product discovery.

Experimental procedures

Sequence database expansion

Experimentally validated PKS BGCs were selected from MIBiG (https://mibig.secondarymetabolites.org/) and published reports. Sequences were extracted from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and KS domains identified using annotations from both published literature and MIBiG. *In silico* predictions were made using the NRPS/PKS prediction tool (67). KS domains were further annotated (*e.g.*, type II noniterative) according to experimentally verified functions and information derived from the associated gene or BGC. Metadata including BGC name, BGC type (*e.g.*, PKS, NRPS), MIBiG accession number, PubMed reference, source organism, and example product name and structure were recorded for each BGC. The complete NaPDoS2 amino acid sequence database (KS and C domains in FASTA format), BGC table, and domain metadata can be downloaded from the NaPDoS2 website.

KS sequence alignment, phylogenetic analysis, and classification

Phylogenies generated from the 1877 KS database sequences (1417 new and 460 existing) were used to establish class and subclass assignments based on correspondence between functional annotation and tree topology. Sequences were aligned using MAFFT (v 7.017) (68) as implemented in Geneious (v 6.1.8) using the FFT-NS-i × 1000 alignment algorithm, BLOSUM62 scoring matrix, and defaults for both gap open penalty and offset value (1.53 and 0.123, respectively). The alignments were trimmed, exported to PHYLIP, and phylogenies generated using the PhyML online tool (http://www.atgc-montpellier.fr/phyml/) with smart model selection enabled (69).

The type II aromatic PKS subclasses were established using a concatenated alignment of the KSα and KSβ subunits and phylogenies generated using the methods described above. Subclasses were delineated based on phylogenetic groupings and features of the core polyketide structure including the length of the poly-β-ketoacyl intermediate, the carbon position of the first ring cyclization, and the type of starter unit. The accuracy of select NaPDoS2 KS classifications from the use case analyses were assessed by analyzing the associated contig with antiSMASH 6.0 (1), the NRPS/PKS prediction tool (67), and transATor (7) (Fig. S6).

KS reference trees

A subset of 414 KS sequences representative of all class and subclass assignments were chosen to generate a KS reference tree (Figs. 1 and S7). Sequences were aligned using MAFFT (68) (v 7.407, FFT-NS-i × 1000 alignment algorithm, BLOSUM62 scoring matrix, default gap open penalty = 1.53, and offset value = 0.123) and trimmed using trimAl (70) (1.4.1 with automatic configuration). The trimmed alignment was used to estimate a maximum likelihood tree using FastTree (71) (v 2.1.11, LG+G model of evolution) with 1000 bootstraps. Booster (72) (v 0.3.1) was used to estimate support as transfer bootstrap expectation values. These programs were implemented in ngphylogeny.fr (71) and run locally as a docker image. Two type II KS phylogenies were generated following the same steps. One comprising all 201 type II KS sequences in the database, eight FAS sequences, and three thiolase outgroup sequences (Figs. 3 and S8). The second used concatenated KSα and KSβ subunits from 59 type II aromatic BGCs (Figs. 4 and S9). Trees were visualized using TreeViewer (https://treeviewer.org/) and annotations added using Adobe Illustrator.

The NaPDoS2 workflow

Database sequences and associated metadata are stored in a back-end MySQL database linked to the NaPDoS2 web portal through CGI-scripting as previously described (8). The transeq tool from the EMBOSS package (v 6.6) is used for 6-frame translations of nucleic acid queries (73). BLAST queries of amino acid sequences are performed using DIAMOND v 0.9.29 (19). Multiple sequence alignments are obtained with MUSCLE v 3.8 (74) using the profile alignment feature to merge query sequences with previously aligned database sequences. Phylogenetic trees are constructed from trimmed amino acid sequences using FastTree v 2.2.1 (75). Graphical tree depictions are generated using Newick Utilities v 1.5.0.

Performance testing

The 1877 full-length amino acid sequences in the NaPDoS2 KS database were clustered at 50% identity using CD-HIT v 4.7 (76), yielding 213 nonredundant positive controls. Using the CD-Search (77) function of the curated NCBI Conserved Domain Database (78), negative controls were selected from subfamilies within the condensing enzyme superfamily (cl09938) that are functionally related to type I and II KS domains but phylogentically distinct from the positive control sequences. An additional 49 sequences (52) and 14 KSs from type III PKSs in the MIBiG 2.0 repository (22) were added for a total of 697 sequences (Table S3), which were also clustered at 50% identity using CD-HIT to obtain 308 nonredundant negative controls.

Cross-validation and receiver operating characteristic curves

Domain detection sensitivity and specificity were determined using BLASTP searches of full-length positive and negative control sequences against the NaPDoS and NaPDoS2 databases, excluding self-matches, to generate leave-one-out cross-validation values. Receiver operating characteristic (ROC) curves were constructed and AUC values calculated from these results using easyROC v 1.3 (79).

EasyROC data output tables were used to identify potential cutoff points based on the most restrictive cross-validation e-value at which 100% of true positives were detected, thus maximizing sensitivity with the minimum possible number of false positives.

KS detection and classification accuracy

A custom perl script (sequence_subdivider.pl, available at https://github.com/spodell/NaPDoS2_website) was used to subdivide the 213 positive control KS sequences into test sets containing overlapping subsequences of 30, 50, 100, or 200 amino acids, each offset by a 10 amino acid sliding window start site. These size-selected test sets, which contained 8555, 8129, 7064, and 4934 subsequences, respectively, were analyzed using the NaPDoS2 workflow to assess the effects of query size on classification accuracy. Accuracy evaluations for each test set were based on the percentage of subsequences whose best nonself, BLASTP match had the same NaPDoS2 classification as the original, full-length sequence from which it was derived.

Application use cases

Accession information for all sequences and datasets analyzed is provided in Table S3. All analyses used the following default settings unless noted otherwise: NaPDoS version 1: domain detection: Hidden Markov Model 1e-5, 200aa minimum alignment length, pathway assignment: e-value cutoff of 1e-5. NaPDoS2: e-value cutoff 1e-8 and 200aa minimum alignment length. Sequence files containing >500,000 sequences or larger than 500 MB were split into smaller subunits using a custom perl script (serialize_seqs.pl, available at https://github.com/spodell/NaPDoS2_website).

Genomes and metagenomes

Salinispora genome protein sequences (53) downloaded from NCBI and JGI IMG/MER (80) were concatenated into a single FASTA file for NaPDoS2 analysis. Fungal genome protein sequences were downloaded from NCBI; fungal PKS BGCs were extracted from the MIBiG 2.0 repository using the query "Kingdom: Fungi AND BGC type: pks". Coding sequences and predicted proteins for *E. chlorotica* were downloaded from NCBI (56). Trimmed *E. chlorotica* KS sequences generated by NaPDoS2 were aligned with previously published EcPKS1, EcPKS2, and EcFAS sequences (57) and used to construct a phylogenetic tree with the closest NaPDoS2 database hits. Marine sediment metagenomes were selected from the Paired Omics Data Platform (60) and downloaded from NCBI SRA.

Amplicons

KS amplicon sequences were analyzed using minimum alignment lengths of 50aa in NaPDoS2 unless otherwise noted. Sequence accession and dataset references are listed in Table S3. Random subsets of query sequences were obtained using custom perl scripts (get_seq_info.pl, randomize_lines.pl, serialize_large_list.pl, and getseq_multiple.pl, available at https://github.com/spodell/NaPDoS2_website).

Data availability

Relevant alignment files, phylogenetic tree files, webtool documentation file, example tutorials, sequence files, and other supporting information can be found on the corresponding OSF project page: https://osf.io/uzhcp/. The code used to construct version two of the Natural Product Domain Seeker website can be found on the corresponding Github repository: https://github.com/spodell/NaPDoS2_website. All accession information and dataset references can be found in Table S3 (separate Microsoft Excel file).

Supporting information—Additional figures and tables referred to in the text (1, 7, 8, 22, 52, 53, 56, 57, 60–65, 67, 74, 75, 78, 81–85). Figs. S1–S13: NaPDoS2 workflow and website screenshots, updated database statistics and classification validation, positive/negative control sequence selection, classification category accuracy validation, phylogenetic context of amplicon dataset analysis, and expanded KS phylogenetic trees.
Tables S1–S12: NaPDoS2 dataset analysis speed comparisons, classification category taxa, application use case analyses of genome, metagenome, and amplicon datasets, and accession information for all analyzed sequences and datasets. Table S3 (Microsoft Excel file): all dataset and sequence accession information.

Acknowledgments—The authors acknowledge James Wang and Ghulam Mustafa for assistance in identifying sequences added to the database. We also thank Dulce Guillén-Matus and Jeong Sang Yi for providing useful feedback on beta-versions of the webtool.

Author contributions—L. J. K., S. P., N. Z., A. C. L., and P. R. J conceptualization; L. J. K., K. E. C., A. M. D., S. P., N. Z., and A. C. L. methodology; L. J. K., K. E. C., A. M. D., S. P., H. W. S., B. S. M., P. R. J., and A. C. L. formal analyses; L. J. K., K. E. C., S. P., and A. M. D. data curation, L. J. K. writing-original draft; S. P., K. E. C. validation; S. P. software; S. P., K. E. C., A. M. D., H. W. S., B. S. M., E. E. A., P. R. J., and N. Z. writing–review and editing; B. S. M. and P. R. J. funding acquisition.

Funding and additional information—This research was supported by the National Institutes of Health (grant no. 5R01GM085770 to P. R. J. and B. S. M.), the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-1650112 to K. E. C. and A. M. D.; grant no. DGE-2038238 to H. W. S.), and the National Science Foundation Division of Molecular and Cellular Biosciences (grant MCB-1149552 to E. E. A.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: ACP, acyl carrier protein; BGC, biosynthetic gene cluster; C, condensation; ECH, enoyl-CoA hydratase; ER, enoyl reductase; FAS, fatty acid synthase; HR, highly reducing; KS, ketosynthase; NaPDoS, Natural Product Domain Seeker; NRPS, nonribosomal peptide synthetase; NR,

nonreducing; PKS, polyketide synthase; PUFA, polyunsaturated fatty acid; PTM, polycyclic tetramate macrolactam; *trans*-AT, *trans*-acyl transferase.

References

1. Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., *et al.* (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucl. Acids Res.* **1**, W29–W35
2. Skinnider, M. A., Johnston, C. W., Gunabalasingam, M., Merwin, N. J., Kieliszek, A. M., MacLellan, R. J., *et al.* (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Comm.* **11**, 6058
3. Medema, M. H. (2021) The year 2020 in natural product bioinformatics: an overview of the latest tools and databases. *Nat. Prod. Rep.* **38**, 301–306
4. Medema, M. H., de Rond, T., and Moore, B. S. (2021) Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571
5. Mungan, M. D., Alanjary, M., Blin, K., Weber, T., Medema, M. H., and Ziemert, N. (2020) Arts 2.0: feature updates and expansion of the antibiotic resistant target seeker for comparative genome mining. *Nucl. Acids Res.* **48**, W546–W552
6. Reddy, B. V. B., Milshteyn, A., Charlop-Powers, Z., and Brady, S. F. (2014) eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chem. Biol.* **21**, 1023–1033
7. Helfrich, E. J., Ueoka, R., Dolev, A., Rust, M., Meoded, R. A., Bhushan, A., *et al.* (2019) Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821
8. Ziemert, N., Podell, S., Penn, K., Badger, J. H., Allen, E., and Jensen, P. R. (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **7**, e34064
9. Hertweck, C. (2009) The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed.* **48**, 4688–4716
10. Schöner, T. A., Gassel, S., Osawa, A., Tobias, N. J., Okuno, Y., Sakakibara, Y., *et al.* (2016) Aryl polyenes, a highly abundant class of bacterial natural products, are functionally related to antioxidative carotenoids. *ChemBioChem* **17**, 247–253
11. Wietz, M., Duncan, K., Patin, N., and Jensen, P. (2013) Antagonistic interactions mediated by marine bacteria: the role of small molecules. *J. Chem. Ecol.* **39**, 879–891
12. Fischbach, M. A., and Walsh, C. T. (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468–3496
13. Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr. Opin. Chem. Biol.* **7**, 285–295
14. Lohman, J. R., Ma, M., Osipiuk, J., Nocek, B., Kim, Y., Chang, C., *et al.* (2015) Structural and evolutionary relationships of "AT-less" type I polyketide synthase ketosynthases. *Proc. Nat. Acad. Sci. U. S. A.* **112**, 12693–12698
15. Grammbitter, G. L., Schmalhofer, M., Karimi, K., Shi, Y.-M., Schoner, T. A., Tobias, N. J., *et al.* (2019) An uncommon type II PKS catalyzes biosynthesis of aryl polyene pigments. *J. Amer. Chem. Soc.* **141**, 16615–16623
16. Metsa-Ketela, M., Halo, L., Munukka, E., Hakala, J., Mantsala, P., and Ylihonko, K. (2002) Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various *Streptomyces* species. *Appl. Environ. Microbiol.* **68**, 4472–4479
17. Jenke-Kodama, H., Sandmann, A., Muller, R., and Dittmann, E. (2005) Evolutionary implications of bacterial polyketide synthases. *Mol. Biol. Evol.* **22**, 2027–2039
18. Moffitt, M. C., and Neilan, B. A. (2003) Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. *J. Mol. Evol.* **56**, 446–457
19. Buchfink, B., Xie, C., and Huson, D. H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Met.* **12**, 59–60
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
21. Sigrist, R., Luhavaya, H., McKinnie, S. M., Ferreira da Silva, A., Jurberg, I. D., Moore, B. S., *et al.* (2020) Nonlinear biosynthetic assembly of alpiniamide by a hybrid cis/trans-AT PKS-NRPS. *ACS Chem. Biol.* **15**, 1067–1077
22. Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J., *et al.* (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucl. Acids Res.* **48**, D454–D458
23. Feng, L., Gordon, M. T., Liu, Y., Basso, K. B., and Butcher, R. A. (2021) Mapping the biosynthetic pathway of a hybrid polyketide-nonribosomal peptide in a metazoan. *Nat. Comm.* **12**, 4912
24. Austin, M. B., Saito, T., Bowman, M. E., Haydock, S., Kato, A., Moore, B. S., *et al.* (2006) Biosynthesis of *Dictyostelium discoideum* differentiation-inducing factor by a hybrid type I fatty acid–type III polyketide synthase. *Nat. Chem. Biol.* **2**, 494–502
25. Coates, R. C., Podell, S., Korobeynikov, A., Lapidus, A., Pevzner, P., Sherman, D. H., *et al.* (2014) Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PLoS One* **9**, e85140
26. Nakamura, H., Hamer, H. A., Sirasani, G., and Balskus, E. P. (2012) Cylindrocyclophane biosynthesis involves functionalization of an unactivated carbon center. *J. Am. Chem. Soc.* **134**, 18518–18521
27. Walker, P., Weir, A., Willis, C., and Crump, M. (2021) Polyketide β-branching: diversity, mechanism and selectivity. *Nat. Prod. Rep.* **38**, 723–756
28. Herbst, D. A., Townsend, C. A., and Maier, T. (2018) The architectures of iterative type I PKS and FAS. *Nat. Prod. Rep.* **35**, 1046–1069
29. Chen, H., and Du, L. (2016) Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. *Appl. Microbiol. Biotechnol.* **100**, 541–557
30. Metz, J. G., Roessler, P., Facciotti, D., Levering, C., Dittrich, F., Lassner, M., *et al.* (2001) Production of polyunsaturated fatty acids by polyketide synthases in both prokaryotes and eukaryotes. *Science* **293**, 290–293
31. Cao, S., Blodgett, J. A., and Clardy, J. J. O. L. (2010) Targeted discovery of polycyclic tetramate macrolactams from an environmental *Streptomyces* strain. *Org. Lett.* **12**, 4652–4654
32. Gallo, A., Ferrara, M., and Perrone, G. (2013) Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins* **5**, 717–742
33. Chooi, Y.-H., and Tang, Y. (2012) Navigating the fungal polyketide chemical space: from genes to molecules. *J. Org. Chem.* **77**, 9933–9953
34. Schmitt, I., and Lumbsch, H. T. (2009) Ancient horizontal gene transfer from bacteria enhances biosynthetic capabilities of fungi. *PLoS One* **4**, e4437
35. Nguyen, T., Ishida, K., Jenke-Kodama, H., Dittmann, E., Gurgui, C., Hochmuth, T., *et al.* (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **26**, 225–233
36. Hertweck, C. (2015) Decoding and reprogramming complex polyketide assembly lines: prospects for synthetic biology. *Trends Biochem. Sci.* **40**, 189–199
37. Bretschneider, T., Heim, J. B., Heine, D., Winkler, R., Busch, B., Kusebauch, B., *et al.* (2013) Vinylogous chain branching catalysed by a dedicated polyketide synthase module. *Nature* **502**, 124–128
38. Chen, A., Re, R. N., and Burkart, M. D. (2018) Type II fatty acid and polyketide synthases: deciphering protein–protein and protein–substrate interactions. *Nat. Prod. Rep.* **35**, 1029–1045
39. Hillenmeyer, M. E., Vandova, G. A., Berlew, E. E., and Charkoudian, L. K. (2015) Evolution of chemical diversity by coordinated gene swaps in type II polyketide gene clusters. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13952–13957
40. Tang, Y., Tsai, S.-C., and Khosla, C. (2003) Polyketide chain length control by chain length factor. *J. Amer. Chem. Soc.* **125**, 12708–12709
41. Komaki, H., and Harayama, S. (2006) Sequence diversity of type-II polyketide synthase genes in Streptomyces. *Actinomycetologica* **20**, 42–48
42. Kim, J., and Yi, G.-S. (2012) PKMiner: a database for exploring type II polyketide synthases. *BMC Microbiol.* **12**, 169

43. Villebro, R., Shaw, S., Blin, K., and Weber, T. (2019) Sequence-based classification of type II polyketide synthase biosynthetic gene clusters for antiSMASH. *J. Ind. Microbiol. Biotechnol.* **46**, 469–475
44. Kwon, H.-J., Smith, W. C., Scharon, A. J., Hwang, S. H., Kurth, M. J., and Shen, B. (2002) CO bond formation by polyketide synthases. *Science* **297**, 1327–1330
45. Walczak, R. J., Woo, A. J., Strohl, W. R., and Priestley, N. D. (2000) Nonactin biosynthesis: the potential nonactin biosynthesis gene cluster contains type II polyketide synthase-like genes. *FEMS Microbiol. Lett.* **183**, 171–175
46. Rebets, Y., Brötz, E., Manderscheid, N., Tokovenko, B., Myronovskyi, M., Metz, P., *et al.* (2015) Insights into the pamamycin biosynthesis. *Angew. Chem. Intern. Ed.* **54**, 2280–2284
47. Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Brown, L. C. W., Mavrommatis, K., *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421
48. Du, D., Katsuyama, Y., Horiuchi, M., Fushinobu, S., Chen, A., Davis, T. D., *et al.* (2020) Structural basis for selectivity in a highly reducing type II polyketide synthase. *Nat. Chem. Biol.* **16**, 776–782
49. Lee, W. C., Choi, S., Jang, A., Yeon, J., Hwang, E., and Kim, Y. (2021) Structural basis of the complementary activity of two ketosynthases in aryl polyene biosynthesis. *Sci. Rep.* **11**, 1–10
50. Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York
51. Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recog. Lett.* **27**, 861–874
52. Jiang, C., Kim, S. Y., and Suh, D.-Y. (2008) Divergent evolution of the thiolase superfamily and chalcone synthase family. *Mol. Phylogenet. Evol.* **49**, 691–701
53. Millán-Aguiñaga, N., Chavarria, K. L., Ugalde, J. A., Letzel, A.-C., Rouse, G. W., and Jensen, P. R. (2017) Phylogenomic insight into *Salinispora* (bacteria, Actinobacteria) species designations. *Sci. Rep.* **7**, 3564
54. Zhang, G., Zhang, W., Saha, S., and Zhang, C. (2016) Recent advances in discovery, biosynthesis and genome mining of medicinally relevant polycyclic tetramate macrolactams. *Curr. Top. Med. Chem.* **16**, 1727–1739
55. Almeida, H., Tsang, A., and Diallo, A. B. (2019) Supporting supervised learning in fungal Biosynthetic Gene Cluster discovery: new benchmark datasets. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1280–1287)
56. Cai, H., Li, Q., Fang, X., Li, J., Curtis, N. E., Altenburger, A., *et al.* (2019) A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. *Sci. Data* **6**, 1–13
57. Torres, J. P., Lin, Z., Winter, J. M., Krug, P. J., and Schmidt, E. W. (2020) Animal biosynthesis of complex polyketides in a photosynthetic partnership. *Nat. Comm.* **11**, 190022
58. Beedessee, G., Hisata, K., Roy, M. C., Satoh, N., and Shoguchi, E. (2015) Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, *Symbiodinium minutum*. *BMC Genomics* **16**, 1–11
59. Oliver, A., Podell, S., Pinowska, A., Traller, J. C., Smith, S. R., McClure, R., *et al.* (2021) Diploid genomic architecture of *Nitzschia inconspicua*, an elite biomass production diatom. *Sci. Rep.* **11**, 1–14
60. Schorn, M. A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D. D., Aksenov, A. A., *et al.* (2021) A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368
61. Wawrik, B., Kerkhof, L., Zylstra, G. J., and Kukor, J. J. (2005) Identification of unique type II polyketide synthase genes in soil. *Appl. Environ. Microbiol.* **71**, 2232–2238
62. Owen, J. G., Reddy, B. V. B., Ternei, M. A., Charlop-Powers, Z., Calle, P. Y., Kim, J. H., *et al.* (2013) Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11797–11802
63. Borsetto, C., Amos, G. C., da Rocha, U. N., Mitchell, A. L., Finn, R. D., Laidi, R. F., *et al.* (2019) Microbial community drivers of PK/NRP gene diversity in selected global soils. *Microbiome* **7**, 1–11
64. Elfeki, M., Alanjary, M., Green, S. J., Ziemert, N., and Murphy, B. T. (2018) Assessing the efficiency of cultivation techniques to recover natural product biosynthetic gene populations from sediment. *ACS Chem. Biol.* **13**, 2074–2081
65. Gumerov, V. M., and Zhulin, I. B. (2020) Trend: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucl. Acids Res.* **48**, W72–W76
66. Bauman, K. D., Shende, V. V., Chen, P. Y.-T., Trivella, D. B., Gulder, T. A., Vellalath, S., *et al.* (2022) Enzymatic assembly of the salinosporamide γ-lactam-β-lactone anticancer warhead. *Nat. Chem. Biol.* **18**, 538–546
67. Bachmann, B. O., and Ravel, J. (2009) Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Met. Enzymol.* **458**, 181–217
68. Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780
69. Lefort, V., Longueville, J.-E., and Gascuel, O. (2017) Sms: smart model selection in PhyML. *Mol. Biol. Evol.* **34**, 2422–2424
70. Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
71. Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., *et al.* (2019) NGPhylogeny. Fr: new generation phylogenetic services for non-specialists. *Nucl. Acids Res.* **47**, W260–W265
72. Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., *et al.* (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456
73. Rice, P., Longden, I., and Bleasby, A. (2000) Emboss: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277
74. Edgar, R. C. (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113
75. Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490
76. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152
77. Marchler-Bauer, A., and Bryant, S. H. (2004) CD-search: protein domain annotations on the fly. *Nucl. Acids Res.* **32**, W327–W331
78. Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucl. Acids Res.* **48**, D265–D268
79. Goksuluk, D., Korkmaz, S., Zararsiz, G., and Karaagaoglu, A. E. (2016) easyROC: an interactive web-tool for ROC curve analysis using R language environment. *R. J.* **8**, 213
80. Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., *et al.* (2019) IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucl. Acids Res.* **47**, D666–D677
81. Marchler-Bauer, A., Anderson, J. B., Derbyshire, M. K., DeWeese-Scott, C., Gonzales, N. R., Gwadz, M., *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucl. Acids Res.* **35**, 237–240
82. Letunic, I., and Bork, P. (2019) Interactive tree of life (ITOL) v4: recent updates and new developments. *Nucl. Acids Res.* **47**, W256–W259
83. Miller, M. A, Pfeiffer, W, and Schwartz, T. Creating the CIPRES science gateway for inference of large phylogenetic trees. In 2010 Gateway Computing Environments Workshop, GCE 2010; New Orleans, LA, 2010; pp 1–8.
84. Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165
85. Stamatakis, A. (2014) RAxML Version 8: A tool for phylogenetic analysis and post- analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313